Data Mining & Statistical Learning

Statistics 640/444 • Fall 2015 • Tuesdays & Thursdays 4:00 - 5:15pm

Instructor:
Genevera Allen
Office: Duncan Hall 2098.
Office Hour: Thursday 5:15 - 6:15pm
Email: gallen@rice.edu

Stat 444 TA:
John Nagorski
Office: Duncan Hall 1038.
Office Hour: TBD.
Email: john.nagorski@rice.edu

Stat 640 TA:
Frederick Campbell
Office: Duncan Hall 1041.
Office Hour: Wednesday 3:00 - 4:00pm
Email: frederick.campbell@rice.edu

Time & Location:

Class: Tuesdays & Thursdays 4:00 - 5:15pm.	Location: Herzstein Hall 210.
Stat 640 Recitation: TBD.	Location: TBD.
Stat 444 Recitation: TBD.	Location: TBD.

Course Description: This course is a survey of statistical learning methods. The course will cover major statistical learning methods and concepts for both supervised and unsupervised learning. Topics covered include penalized regression and classification, support vector machines, kernel methods, model selection, matrix factorizations, graphical models, clustering, boosting and ensemble learning.

Course Objectives & Outcomes: Students will learn how and when to apply statistical learning techniques, their comparative strengths and weaknesses, and how to critically evaluate the performance of learning algorithms. Students completing this course should be able to (i) apply basic statistical learning methods to build predictive models or perform exploratory analysis, (ii) properly tune and select statistical learning models, (iii) correctly assess model fit and error, and (iv) build an ensemble of learning algorithms.

Stat 640 vs. 444: Students registered in 444 and 640 will have different homework assignments, different midterm exams, and be graded by different criteria for the data mining competition. Stat 444 will focus on applications and the interpretation of statistical learning methods. Stat 640 will additionally focus on how these methods work, how they are implemented, and their mathematical properties.

Prerequisites: Linear Algebra & Scientific Programming in a language such as `R`, `Matlab`, or `Python`. Recommended: Linear Regression & Mathematical Probability and Statistics. Please speak to the instructor if you have any concerns regarding prerequisites.

Recommended Textbooks: Elements of Statistical Learning by Hastie, Tibshirani & Friedman; available online at `http://www-stat.stanford.edu/~tibs/ElemStatLearn`. (640) Statistics for High-Dimensional Data by Buhlmann & van de Geer.
(444) Introduction to Statistical Learning by James, Witten, Tibshirani & Hastie; available online at `http://www-bcf.usc.edu/~gareth/ISL/`.

Course Webpage: `http://www.stat.rice.edu/~gallen/stat640.html`. Please check the course webpage frequently for announcements and homework assignments.

Grading Policy:

Homeworks	25%
Mid-Term Exam	25%
Competition & Reports	45%
Class Participation	5%

Registrar Deadlines:

Friday, September 4	Add Deadline
Friday, October 9	Drop Deadline

Attendance Policy: Student are encouraged, but not required to attend lectures.

Homeworks: There will be four(+1) homework assignments. Approximate dates:

	Homework Assigned	Homework Due
Homework 0	August 25	September 1
Homework I	September 3	September 17
Homework II	September 22	October 8
Homework III	October 15	October 29
Homework IV	November 3	November 24

A hard copy of homeworks are due at 5:15pm and can be turned in at class or to the TA in Duncan Hall on the due date. Late homeworks will NOT be accepted, NO exceptions. Homeworks may be discussed with classmates but must be written and submitted individually.

Midterm Exam: There will be an in-class midterm exam assigned on **November 5**. The exam is open books and open notes. Internet use is prohibited and the exam must be completed individually according to the honor code.

Data Mining Competition: As part of the class, students will compete against each other in a data mining contest. The competition will begin on **Tuesday, September 8** and can be completed in teams of two people. Grades will be based upon two progress reports and a final report (one per team) as well as the contest results. The winning teams for 444 and 640 in terms of prediction accuracy, the team with the most innovative solution (640), and the team with the most interesting finding (444) will automatically receive the highest grades. Further details about the contest along with specific grading criteria will be given in a separate document and discussed in class.

Contest Opens	12:00am September 8
Progress Report I	5:15pm October 1
Progress Report II	5:15pm November 3
Contest Closes	11:59pm November 30
In-Class Contest Presentations	December 1 & 3
Final Report Due	5:00pm December 4

Tentative Schedule: (Also see separate schedule document).

August 25	Introduction & KNN
August 27	MSE & Least Squares
September 1	Ridge, PCA & PLS Regression
September 3	Sparse Regression I
September 8	Sparse Regression II
September 10	Sparse Regression III
September 15	Linear Discriminant Analysis
September 17	LDA & Logistic Regression
September 22	Sparse Classification
September 24	Support Vector Machines I
September 29	SVMs II
October 1	Kernel Methods
October 6	Model Selection I
October 8	Model Selection II
October 15	Matrix Factorizations I
October 20	Matrix Factorizations II
October 22	Clustering I
October 27	Clustering II
October 29	Graphical Models I
November 3	Graphical Models II
November 5	**In-Class Midterm Exam**
November 10	Trees
November 12	Ensemble Learning & Boosting
November 17	Boosting
November 19	Random Forests
November 24	Ensemble Learning
December 1	Competition Presentations
December 3	Competition Presentations

Students with a documented disability requiring accommodations should speak with the instructor during the first two weeks of class.

This syllabus is subject to change with reasonable advance notice by the instructor.